DOCUMENT RESUME

ED 406 444 TM 026 423

AUTHOR Rodriguez, Maximo
TITLE Non-Factorial ANOVA: Test Only Substantive and Interpretable Hypotheses.
PUB DATE Jan 97
NOTE 21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Factor Analysis; *Hypothesis Testing; *Research Design; Test Interpretation
IDENTIFIERS *Experimentwise Error; Type I Errors

ABSTRACT

Factorial analyses differ from nonfactorial analyses in that in the former all possible hypotheses (all possible main effects and interaction effects) are tested regardless of their substantive interest to the researcher and their interpretability, while in the latter, only substantive and interpretable hypotheses are tested. This paper shows the circumstances under which nonfactorial analyses are more appropriate than factorial ones. Balanced factorial analyses maximally inflate the experimentwise error rate, and in so doing, they increase the likelihood of making Type I errors. Hypothetical experiments are used to make the discussion more concrete. It is argued that only substantive and interpretable hypotheses in the design should be tested. Uninterpretable hypotheses increase the probability of making Type I error by increasing the experimentwise error rate. (Contains 2 tables and 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 406 444

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
MAXIMO RODRIGUEZ
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

NON-FACTORIAL ANOVA: TEST ONLY SUBSTANTIVE AND INTERPRETABLE HYPOTHESES

Maximo Rodriguez

Texas A&M University 77843-4272

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January, 1997.

TM026423

BEST COPY AVAILABLE

Abstract

Factorial analyses differ from non-factorial analyses in that in the former all possible hypotheses (all possible main effects and interaction effects) are tested regardless of their substantive interest to the researcher and/or their interpretability, while in the latter only substantive and interpretable hypotheses are tested. In the present paper it is shown how in some cases non-factorial analyses are more appropriate than factorial ones. Hypothetical experiments are utilized to make the discussion more concrete. It is argued that only substantive and interpretable hypotheses in the design should be tested.

Since Cohen's (1968) seminal article in which he argued that ANOVA and ANCOVA are especial cases of multiple regression analysis, criticisms against the application of ANOVA-type methods (ANOVA, ANCOVA, MANOVA, MANCOVA--hereafter labeled OVA methods) have grown stronger. Major criticisms have centered around the categorization of intervally-scaled independent variables in OVA analyses.

As Pedhazur (1982, pp. 453-454) noted,

> Categorization leads to a loss of information, and consequently to a less sensitive analysis ... all subjects within a category are treated alike even though they may have been originally quite different in the continuous variable ... It is this loss of information about the differences between subjects, or the reduction in the variability of the continuous variable, that leads to a reduction in the sensitivity of the analysis, not to mention the meaningfulness of the results.

Pedhazur and Pedhazur-Schmelkin (1991, p. 539) argued that categorization of continuous variables has even more harmful effects. First, the nature of the variable changes, as it is generally treated as if it were a categorical variable, not as a continuous variable that has been categorized ... As a result of the change in the nature of variable, the very idea of trends (e.g., linear, quadratic) in the data is precluded. Second, categorization of continuous variable in nonexperimental research and casting the design in an ANOVA format tends to create the false impression that a nonexperimental design has thereby been transformed into an experimental design, or at the very least, into something close approximating it.

Discarding variance is not generally regarded as good research practice (Thompson, 1988). As kerlinger (1986, p. 558) pointed out, "variance is the 'stuff' on which all analysis is based." Of course, as Haase and Thompson (1992, p. 4) stated, "Anova does remain a useful tool when the independent variables are inherently nominal (e.g., dichotomies or trichotomies such as assignment to experimental condition and gender)."

Despite the criticisms against OVA methods, empirical studies of behavioral research practice (Edgington, 1974; Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980) indicate that these methods are still very popular among social scientists. Oftentimes, in their attempts to identify the variables contributing to a given phenomenon, behavioral researchers design experiments in which the focus of attention is on the effect of one independent variable or factor on some dependent variable (a single-factor design). However, in some instances researchers become more interested in assessing the effects of two or more independent variables on a single dependent variable. This is typically accomplished through a factorial design. In both cases, those researchers resort to the classical ANOVA--one-way ANOVA for the first kind of design and factorial multi-way ANOVA (also called factorial analysis) for the second. Factorial analyses differ from non-factorial analyses in that in the former all hypotheses (all possible main effects and interaction effects) are tested regardless of their meaningfulness and/or interpretability, while in the latter, only those hypotheses of interest to the researcher are tested.

The purpose of the present paper is to show how in some cases non-factorial analyses are more appropriate than factorial ones. It is argued that only substantive and interpretable hypotheses in the design should be tested. Hypothetical experiments are utilized to make the discussion more concrete. A brief description of factorial designs is provided to establish a context for the discussion.

Factorial Designs

Factorial designs permit the manipulation of more than one independent variable in the same experiment. The arrangement of the treatment conditions is such that information can be obtained about the influence of the independent variables considered separately and about how the variables combined to influence behavior (Keppel, 1991, p. 19).

As Keppel and Zedeck (1989) noted, factorial designs are usually described in regard to the number of levels associated with the independent variables. Thus, a 2 x 3 factorial design clearly specifies that two independent variables have been manipulated factorially, one with two levels and the other with three levels, and that the total number of treatment conditions or cells is six.

Factorial designs may consist of more than two factors, each comprised of any number of levels. For example, a 2 x 2 x 3 x 5 is a four-factor design: two factors with 2 levels, one with 3 levels, and one with 5 levels. Designs consisting of more than two factors are referred to as higher-order designs.

To use a concrete example, suppose that a researcher designs a completely randomized 3 x 2 factorial experiment (also called a between-subjects design). By

completely randomized we mean that each subject is randomly assigned to only one of the six treatment conditions; in other types of designs subjects are either exposed to all treatment conditions in a randomized order (a within-subject design) or they are exposed to some, but not all, of the treatment conditions defined by the factorial design (a mixed-factorial design). Let us assume that the factors manipulated concurrently in this hypothetical experiment are: EFL vocabulary teaching methodology (the keyword method, the semantic approach, the keyword-semantic approach), and time (immediate and delay), while the dependent variable is cued recall. Let us also assume an equal number of subjects per cell (a balanced or orthogonal design). Incidentally, only balanced designs are discussed in this paper (for discussion of unbalanced designs, see Hays, 1991; Keppel, 1991; Keppel & Zedeck, 1989; Pedhazur & Pedhazur-Schmelkin, 1991).

In this hypothetical experiment, the six treatment combinations: keyword-immediate, keyword-delay, semantic-immediate, semantic-delay, keyword-semantic-immediate, and keyword-semantic-delay, are specified in the following matrix:

TEACHING METHOD	TIME	
	IMMEDIATE	DELAY
KEYWORD		
SEMANTIC		
KEYWORD-SEMANTIC		

As Keppel (1991, p. 188) explained, a factorial design consists of a set of single-factor experiments. Thus, from our hypothetical experiment, the researcher may create

two single-factor experiments: one may consist of three groups of learners randomly assigned to a different vocabulary teaching method, tested immediately after the presentation of the language material. This single experiment assesses the effects of vocabulary teaching method on cued recall under the condition of immediate recall, and if the manipulation were successful, the researcher would attribute any differences among the groups of learners to the differential effectiveness of the teaching methods. The other experiment would be an exactly duplicate of the first except that learners would be tested some time (e.g., 2 days, or a week, or 10 days) after the presentation of the input material.

This hypothetical factorial design can also be viewed as a set of component-single factor experiments involving the other independent variable, time. In this case, the researcher may create three single-factor experiments: one may consist of two groups of learners randomly assigned to the two time conditions and instructed with the keyword method. The two other experiments would be exact duplicates of the first, except that learners would be taught with the semantic approach in one and with the keyword-semantic in the other. Each component experiment provides information about the effects of time, but for different vocabulary teaching methods.

The results of these component single-factor experiments are called the simple effects of an independent variable. These effects reflect treatment effects associated with one of the independent variable, with the other held constant. Besides simple effects, factorial designs produce two other important pieces of information: main effects and interaction effects. Main effects are referred to as the deviation of a category or level mean from the grand mean, and essentially transform the factorial design into a set of

single-factor experiments, while the interaction effects reflect a comparison of the simple effects.

In our hypothetical 3 x 2 experimental design, a main effect for vocabulary teaching method would mean that there are differences in the effectiveness of these methods regardless of whether cued recall is ascertained immediately after the presentation of the language material versus some time later. On the other hand, a main effect of time would mean that learners' performance on immediate and delayed recall is different regardless of the teaching method used. Finally, an interaction effect would mean that the effect of the teaching methods on learners' cued recall is not constant under the two time conditions. The advantages of factorial designs over single-factor experiments are widely recognized by most researchers, and are briefly discussed here.

Advantages of Factorial Designs

Pedhazur (1982, p. 135) discussed four major advantages of factorial designs:

1) Factorial designs make it possible to determine whether the independent variables interact in their effect on the dependent variable. An independent variable can explain a relatively small proportion of variance of a dependent variable, whereas its interaction with other independent variables may explain a relatively large proportion of the variance. Studying the effects of the independent variables in isolation cannot reveal the interaction between them.

2) Factorial designs afford the researcher greater control, and, consequently more sensitive (i.e., statistically powerful) statistical tests than the statistical tests used in analysis with single variables.

3) Factorial designs are efficient. One can test the separate and combined effects of several variables using the same number of subjects one would have used for separate experiments.

4) In factorial designs the effect of a treatment is studied across different conditions of other treatments. Consequently, generalizations from factorial experiments are broader than generalizations from single-variable experiments. Factorial designs are examples of efficiency, power, and elegance.

Interpretation of Factorial Analysis

Keppel (1991) argued that the test of interaction is usually the logical first step in the analysis of factorial designs. The results of this test influence the analysis of the main effects. For example, if the interaction is statistically significant, less attention is generally paid to the interpretation of the main effects. After all, as Pedhazur and Pedhazur-Schmelkin (1991, p. 514) noted,

> The motivation for studying interactions is to ascertain whether the effects of a given factor vary depending on the levels of the other factor with which they are combined. Having found this to be the case (i.e., that the interaction is statistically significant), it makes little sense to act as if it is not so, which is what the interpretation of main effects amounts to. Instead, differential effects of the various treatment conditions should be studied ... this is accomplished by doing what are referred to as tests of simple main effects.

On the other hand, if the interaction is not statistically significant, or if it is statistically significant but trivial (according to the researcher's judgment), the attention focuses on the detailed analysis of the main effects. If the main effects are statistically significant, post hoc comparisons should then be tested. However, a statistically significant interaction does not mean that absolutely no attention should be paid to main effects. A large main effect, relative to an interaction, indicates that we should consider both the main effect and the interaction when we describe or interpret our data (Keppel, 1991, p. 232).

The Use of Factorial and Non-Factorial Analysis

As said previously, factorial analyses differ from non-factorial ones in that in the former all possible hypotheses are tested regardless of their substantive interest to the researcher and/or their interpretability, while in the latter only substantive and interpretable hypotheses are tested. Although substantive considerations as the guiding principle for hypothesis testing have been strongly recommended by several scholars (Hays, 1981; Keppel, 1991; Keppel & Zedeck, 1989; Pedhazur & Pedhazur-Schmelkin, 1991; Thompson, 1994), many researchers invariably conduct factorial analyses, and frequently end up testing irrelevant omnibus hypotheses or hypotheses they are unable to interpret, as perhaps in a five-way interaction test. As Thompson (1994, p. 10) explained,

> Some researchers always test even omnibus effects that are not of interest because they naively believe that such analyses always increase the probability of detecting statistically significant effects on the omnibus hypotheses that are of interest.

These researchers do not realize that this is not always the case, and that in fact, it is also possible that testing irrelevant omnibus hypotheses can make substantive effects become statistically nonsignificant. We will use our hypothetical 3 x 2 experiment to illustrate both possibilities. Suppose for example, that the researcher is really only interested in testing the interaction omnibus hypothesis.

Table 1. An Example of How Factorial Analysis Can Help Yield Significance for Effects of Interest by Analyzing Even Effects Not of Interest

Source	SOS	df	MS	Fcal	Fcrit	Dec
			Nonfactorial analysis			
Method x Time	**25.00**	2	**12.50**	3.125	3.29	NS
Residual	132.00	33	4.00			
Total	**157.00**	35				
			Factorial Analysis			
Main						
Method	14.00	2	7.00	1.912	3.32	NS
Time	8.00	1	8.00	2.213	4.17	NS
Method x Time	**25.00**	2	**12.50**	3.415	3.32	Rej
Residual	110.00	30	3.66			
Total	**157.00**	35				

Note. Entries in **bold** remain constant.

Table 1 presents the results of two analyses, and shows how the test for the substantive hypothesis yields a statistical nonsignificant result when it is the only hypothesis tested, and how it becomes statistically significant when the omnibus main effect hypotheses are tested. As can be seen from Table 1, the sum of squares (SOS) for the interaction effect remained constant (25) in both analyses. However, the factorial analysis reduced the sum square error by 22 (132-110), and the degrees of freedom error by 3 (33-30), which made the mean square (MS) error smaller (3.66 versus 4.00). A smaller MS error resulted in a larger F calculated value (3.415), slightly greater than the F critical value (3.32).

Table 2 below presents results from the same design hypothetically implemented with different subjects. This Table illustrates how a statistically significant omnibus test may become statistically nonsignificant because a factorial analysis--the default in many statistical packages--was conducted. In this case, testing only the omnibus interaction hypothesis yields a statistically significant result. Nonetheless, no null hypotheses got rejected when the factorial analysis was performed with the same data. As in the previous example, the SOS for the interaction effect was held constant in both analysis. In the factorial analysis the degrees of freedom error were again reduced by 3 (33-30). However, this time the reduction of the SOS error was very small (115.5-112 = 2.5) which in turn made the MS error larger (3.73 versus 3.50). A larger MS error resulted in a smaller F calculated value (3.217 versus 3.429). This F calculated value is smaller than the F critical value (3.32). Incidentally, it is interesting to point out that due to the

reduction of the degrees of freedom error in a factorial analysis, the F critical values for the omnibus tests become larger.

Table 2. An Example of How Factorial Analysis Can Hurt by Yielding Nonsignificance for the Effects of Primary Interest

Source	SOS	df	MS	Fcal	Fcrit	Dec
			Nonfactorial Analysis			
Method x Time	**24.00**	2	**12.00**	3.429	3.29	Rej
Residual	115.50	33	3.50			
Total	**139.50**	35				
			Factorial Analysis			
Main						
Method	2.50	2	1.25	.335	3.32	NS
Time	1.00	1	1.00	.268	4.17	NS
Method x Time	**24.00**	2	**12.00**	3.217	3.32	NS
Residual	112.00	30	3.73			
Total	**139.50**	35				

Note. Entries in **bold** remain constant

Another issue to be considered by users of factorial analyses deals with Type I error. Two Type I error rates have been identified: testwise (TW) error rate, and experimentwise (EW) error rate. TW error rate refers to the probability of making a Type

I error when testing a given hypothesis. EW error rate refers to the probability of making one or more Type I error anywhere in the whole set of hypotheses tested in the study.

In the case of a study in which only one hypothesis is tested, the TW error rate equals the EW error rate. However, when several hypotheses are tested within a single study, the EW error rate will get inflated unless all the hypotheses are perfectly correlated (Thompson, 1994, p. 6). Most researchers are completely unaware that the use of factorial analyses of balanced designs maximally inflate experimentwise (EW) error since (a) the maximum number of tests are conducted, and (b) the omnibus tests are perfectly uncorrelated in balanced designs (Benton, 1991, p. 125).

The formula for computing EW error rate = $[\ 1 - (1 - TW)^{k}\ ]$, where k is the number of hypotheses tested. Thus, in our hypothetical 3 x 2 factorial design in which both main effect omnibus hypotheses and the two-way omnibus interaction are tested at the .05 level, the EW error rate would be about .14. That is .14 equals

$1 - (1 - .05)^3 =$

$1 - (.95)^3 =$

$1 - .8574 = .14.$

As can be seen from our hypothetical example, by conducting a factorial analysis rather than testing only the hypothesis of interest, the researcher increased by almost three times the probability of making a Type I error in testing the omnibus hypotheses. The potential EW error rates in complex multi-way factorial analyses can be extremely high. Very few researchers and even fewer textbook authors consciously recognize that

inflation of EW error rates occurs in classical OVA methods testing omnibus effects prior to the use of unplanned comparisons (Thompson, 1994, p. 9).

Unplanned (also called a posteriori or post hoc or unfocused) multiple comparison test (e.g., Duncan, Scheffe, Tuckey) are among the choices that can be used to isolate means that are significantly different within OVA ways having more than two levels (Thompson, 1994, p. 4). Post hoc or multiple comparisons is a somewhat derogatory term that generally refers to the indiscriminate examination of all possible comparisons to locate significant effects (Keppel & Zedeck, 1989, p. 149). These comparisons are conducted only if omnibus test results are statistically significant. Thus, simple effects are examined only when the interaction is statistically significant; simple comparisons only when a simple effect is statistically significant; and main comparisons only when a main effect is statistically significant.

Keppel (1991, pp. 247-248) argued that in order to deal with the increase of EW error (what he calls "familywise" error), methodologists have introduced a wide variety of adjustment techniques, but that none of these has captured the attention of researchers except, perhaps, a Bonferroni adjustment for simple effects. This correction usually consists of controlling EW error for the entire set of simple effects, which is accomplished by using alpha = .05/b as the significance level for evaluating the simple effects of factor A and alpha = .05/a for the simple effects of factor B, where a and b refer to the number of levels in the factors. He stated however, that current practice in psychological research favors analyses without correction for EW error rate. It should be pointed out that post hoc tests contain their "built-in" correction factors.

We have to keep in mind that adjustments for EW error rate reduce the sensitivity (or power) of the test. In other words, guarding against Type I error increases the probability of making a Type II error (that is, no rejection of a false null hypothesis). That is why planned (also called a priori or focused) comparisons are a better alternative. Since fewer hypotheses are tested, planned comparisons either orthogonal or nonorthogonal have more statistical power than unplanned comparisons.

A final remark regarding factorial analyses deals with the interpretability of the hypotheses tested. As mentioned earlier, interpretability of the hypotheses tested is not a requirement in factorial analyses. In planning an experiment, it is a temptation to throw in many experimental treatments, especially if the data are inexpensive and the experimenter is adventuresome (Hays, 1981, p. 368).

Although higher-order designs may be advantageous to researchers in some respects, the inclusion of a large number independent variables in a study may be also be problematic as these designs carry with them the possibility of statistically significant higher-order interactions, some of which are simply uninterpretable. The description of higher-order interactions typically requires an extremely complicated statement. As Keppel (1991, p. 482) observed,

> With two-way factorial, an interaction indicates that any description of the influence of one of the factors demands consideration of the specific levels represented by the other factor. With a three-way factorial, a significant higher-order interaction implies that any description of one of the two-way interactions

must be made with reference to the specific levels selected for a third factor. Interactions involving four variables require even more complicated descriptions. Now, if it is difficult to merely summarize the pattern of a particular interaction, imagine the problem we will have in explaining these results.

To illustrate the problems associated with the interpretability of higher-order interactions we will expand our hypothetical 3 x 2 factorial design. Suppose, for example, that the researcher decides to make it a lot more complex by including three other independent variables: time of instruction delivery (morning, afternoon), sex and age of the subjects. Let us assume that age is categorized into 3 levels: younger children (6-12 years old); older children (13-19); and adults (20 on). For the sake of illustration, we will disregard the problems generated by categorizing age, a continuous variable. As a result, our original 3 x 2 factorial design became a 3 x 3 x 2 x 2 x 2 factorial design. This five-way factorial produces a total of 26 interactions.

Let us suppose now that such five-way factorial analysis yielded a statistical significant five-way interaction, and some four-way interactions. How will our researcher interpret these results? The researcher will not be able to do it because these kinds of interactions are typically uninterpretable. Thus, what then is the point of testing uninterpretable hypotheses? Testing this kind of hypotheses not only increases the probability of making Type I errors but reduces the power of the statistical analysis.

Summary

In the present paper the use of factorial and non-factorial analysis was discussed. Using hypothetical experimental data it was illustrated how in some situations, factorial analyses may be advantageous to the researcher and how in some other situations, they could be a detriment to the study's outcome. Power issues related to factorial and nonfactorial analyses were also briefly examined. It was claimed that balanced factorial analyses maximally inflate the EW error rate, and in doing so they increase the likelihood of making Type I errors. Additionally, it was claimed that attempts to control for the EW error rate reduces the power of the statistical analyses. Finally, it was argued that it is nonsensical to test uninterpretable hypotheses for they do not convey any substantive information. On the contrary, they increase the probability of making Type I error by increasing the EW error rate.

References

Benton, R. (1991). Statistical power considerations in ANOVA. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (vol. 1, pp. 119-132). Greenwhich, CT: JAI Press.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Edgington, E. (1974). A new tabulation of statistical procedures used in APA journals. American Psychologists, 29, 25-26.

Elmore, P., & Woehlke, P. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17 (9), 19-20.

Goodwin, L., & Goodwin, W. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14 (2), 5-11.

Haase, T., & Thompson, B. (1992, January). The homogeneity of variance assumption in ANOVA: What it is and what it is required. Paper presented at the annual meeting of the Southwest Educational Research Association.

Hays, W. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ.: Prentice Hall.

Keppel, G., & Zedeck, S. (1989). Data analysis for research designs. New York: W.H. Freeman.

Kerlinger, F. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Pedhazur, E. (1982). Multiple regression in behavioral research: Explanation and prediction. New York: Holt, Rinehart and Winston.

Pedhazur, E., & Pedhazur-Schmelkin, L. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ.: Lawrence Erlbaum Associates Publishers.

Thompson, B. (1988). Discard variance: A cardinal sin in research. Measurement and evaluation in counseling and development, 21, 3-4.

Tompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neoclassical perspective. In B. Thompson (Ed.), Advances in Social Science Methodology (vol. 3, 3-27). Greenwhich, CT: JAI Press.

Willson, V. (1980). research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9, 5-10.